MAXIMUM LIKELIHOOD ESTIMATION OF THE NUMBER OF NUCLEOTIDE SUBSTITUTIONS FROM RESTRICTION SITES DATA

MASATOSHI NEI AND FUMIO TAJIMA

Center for Demographic and Population Genetics, The University of Texas at Houston, Houston, Texas 77025

Manuscript received February 21, 1983
Revised copy accepted May 13, 1983

ABSTRACT

A simple method of the maximum likelihood estimation of the number of nucleotide substitutions is presented for the case where restriction sites data from many different restriction enzymes are available. An iteration method, based on nucleotide counting, is also developed. This method is simpler than the maximum likelihood method but gives the same estimate. A formula for computing the variance of a maximum likelihood estimate is also presented.

THE number of nucleotide substitutions between a pair of homologous DNAs can be estimated from data on restriction enzyme cleavage sites (UPHOLT 1977; NEI and LI 1979; KAPLAN and LANGLEY 1979; GOTOH *et al.* 1979). When all the restriction enzymes used have the same number of nucleotides in their recognition sequence, the number of nucleotide differences per site can be estimated by a simple formula. However, when restriction enzymes with different numbers of recognition nucleotides are used, a rather complicated procedure of maximum likelihood estimation is used (KAPLAN and LANGLEY 1979; KAPLAN and RISKO 1981). GOTOH *et al.* (1979) used a relatively simple maximum likelihood method, but their formulation does not seem to be accurate. In the following we would like to present a simple method of maximum likelihood estimation. We shall also present a method for estimating the number of nucleotide substitutions by means of nucleotide counting and show that this method gives the maximum likelihood estimate.

MAXIMUM LIKELIHOOD METHOD

We assume that the four types of nucleotides (T, C, A, G) are randomly arranged in the DNA sequence under investigation, and the evolutionary change of DNA sequence occurs solely by random nucleotide substitution. (See DISCUSSION for the effect of violation of this assumption.) In this case the expected number of restriction sites for a restriction enzyme with a recognition sequence of r nucleotides (usually $r = 4$ or 6) is given by $m_T a$, where m_T is the total number of nucleotides and a is the probability that a sequence of r nucleotides in the DNA is a restriction site. In general, $a = g_1^{r_1} g_2^{r_2} g_3^{r_3} g_4^{r_4}$, where

g_1, g_2, g_3 and g_4 are the frequencies of nucleotides T, C, A and G, respectively, in the 5′-3′ strand of DNA, and r_1, r_2, r_3 and r_4 are the numbers of T, C, A and G in the recognition sequence, respectively ($\Sigma r_j = r$). For example, the recognition sequence of *Eco*RI is GAATTC, so that $r_1 = 2$, $r_2 = 1$, $r_3 = 2$ and $r_4 = 1$. When a restriction enzyme identifies more than one type of recognition sequences (*e.g.*, *Hae*I), a is given by a somewhat different formula as will be discussed later. Usually, a is much smaller than 1.

We consider two DNA sequences (X and Y) that diverged t years (or generations) ago and compare all possible restriction sites of the two sequences. We note that, in a circular DNA of m_T nucleotides, there are m_T possible restriction sites (NEI and LI 1979). In a linear DNA the possible number of restriction sites is $m_T - r + 1$. However, m_T is usually much larger than r, so that the possible number is again approximately m_T. In the comparison of restriction sites between two DNA sequences, there are four different cases. A sequence of r nucleotides at a particular location of the DNA can be a restriction site (1) for both X and Y, (2) for X but not for Y, (3) for Y but not for X and (4) for neither of X and Y. Let m_X and m_Y be the numbers of restriction sites for DNA sequences X and Y, respectively, and m_{XY} be the number of restriction sites shared by X and Y. The numbers of observations for these four events are then given by m_{XY}, $m_X - m_{XY}$, $m_Y - m_{XY}$ and $m_T - m_X - m_Y + m_{XY}$, respectively.

Let us now derive the probabilities of these four events, considering restriction enzymes with a unique recognition sequence. Let w_i be the probability that a sequence of r nucleotides at a location of the DNA is different from the recognition sequence by i nucleotides, p be the probability that a restriction site in a DNA sequence disappears during t years and q_i be the probability that a site (a sequence of r nucleotides) which was originally different from the recognition sequence by i nucleotides becomes a restriction site during t years. Since the expected number of restriction sites remains constant over time, we have the relationship $ap = \Sigma_{i=1}^{r} w_i q_i$. The probability that a sequence of r nucleotides in the DNA is a restriction site for both X and Y is then given by $a(1-p)^2 + \Sigma_i w_i q_i^2 = a[(1-p)^2 + \Sigma_i w_i q_i^2/a]$. This may be written as aS, where $S = (1-p)^2 + \Sigma_i w_i q_i^2/a$. The probability that the same nucleotide sequence is a restriction site for X but not for Y is $ap(1-p) + \Sigma_i w_i q_i(1-q_i) = a[1 - \{(1-p)^2 + \Sigma_i w_i q_i^2/a\}] = a(1-S)$. The probability of the third event is the same as that for the second event. The probability that the sequence is not a restriction site for both X and Y is $ap^2 + \Sigma_i w_i(1-q_i)^2 = 1 - a[2 - (1-p)^2 - \Sigma_i w_i q_i^2/a] = 1 - a(2-S)$ since $\Sigma_i^r w_i = 1 - a$.

In these probabilities S may be written as $(1-\pi)^r$, where π is the probability that sequences X and Y have different nucleotides at a given nucleotide position. This π is related to the expected number of nucleotide substitutions per site (δ) by

$$\pi = \frac{3}{4}\left[1 - e^{-(4\delta/3)}\right] \qquad (1)$$

(JUKES and CANTOR 1969). (This equation is dependent on the assumption that

the substitution rates among the four different nucleotides are equal, but as long as $\delta < 0.5$, it holds approximately even if this assumption is violated; see DISCUSSION.) If the average rate of nucleotide substitution per site per year is λ, δ is given by

$$\delta = 2\lambda t. \tag{2}$$

Therefore, it is possible to estimate δ if we know S. For restriction enzymes with multiple recognition sequences, $S = (1 - \pi)^r$ does not hold, but if we redefine r as given in a later section, it applies approximately.

We also note that S may be approximated by $(1 - p)^2$, since $\Sigma_i w_i q_i^2/a$ is usually much smaller than $(1 - p)^2$ (NEI and LI 1979). This approximation amounts to neglecting shared restriction sites that have newly arisen by independent mutations in X and Y. In practice, an even more approximate formula for S, *i.e.*, $S = e^{-r\delta}$, may be used as long as $\delta < 0.25$ (NEI and LI 1979; LI 1981; KAPLAN and RISKO 1981). In this case $p = 1 - e^{-r\lambda t}$. In the following we use this relationship unless it is mentioned otherwise.

Under our assumptions the likelihood of having the observed values of m_X, m_Y and m_{XY} is given by

$$\begin{aligned} L = C(aS)^{m_{XY}}[a(1 - S)]^{m_X + m_Y - 2m_{XY}} \\ \times [1 - a(2 - S)]^{m_T - m_X - m_Y + m_{XY}}, \end{aligned} \tag{3}$$

where C is a constant. This is equivalent to KAPLAN and RISKO'S (1981) more complicated equation, which was derived by a different method.

In the above formulation we considered only one restriction enzyme or one type of restriction enzymes with the same r value. If k different types of enzymes are used, the total likelihood is given by the product of the likelihoods for individual types of enzymes. That is,

$$\begin{aligned} L = C \prod_{i=1}^{k} (a_i S_i)^{m_{XYi}}[a_i(1 - S_i)]^{m_{Xi} + m_{Yi} - 2m_{XYi}} \\ \times [1 - a_i(2 - S_i)]^{m_T - m_{Xi} - m_{Yi} + m_{XYi}}, \end{aligned} \tag{4}$$

where i refers to the ith type of enzymes. Previously, we defined a_i in terms of nucleotide frequencies. In practice, we do not know nucleotide frequencies, but a_i can be estimated simultaneously with δ. The maximum likelihood estimates of δ and a_i are given by solving the following equations.

$$\frac{\partial \ln L}{\partial \delta} = -\Sigma_i r_i S_i \left[\frac{m_{XYi} - (m_{Xi} + m_{Yi} - m_{XYi})S_i}{S_i(1 - S_i)} + \frac{(m_T - m_{Xi} - m_{Yi} + m_{XYi})a_i}{1 - a_i(2 - S_i)} \right] \tag{5}$$

$$= 0$$

$$= \Sigma_i \frac{r_i[2m_{XYi} - (m_{Xi} + m_{Yi})S_i]}{(1 - S_i)(2 - S_i)} = 0, \tag{5a}$$

$$\frac{\partial \ln L}{\partial a_i} = \frac{1}{a_i} \left[m_{Xi} + m_{Yi} - m_{XYi} - \frac{a_i(m_T - m_{Xi} - m_{Yi} + m_{XYi})(2 - S_i)}{1 - a_i(2 - S_i)} \right] = 0. \tag{6}$$

From (6) we obtain

$$a_i = \frac{m_{Xi} + m_{Yi} - m_{XYi}}{m\ (2 - S)}. \tag{7}$$

Equation (5a) has been obtained by putting (7) into (5). It is noted that (5a) is quite different from GOTOH *et al.*'s (1979) equivalent formula. This difference occurred because they did not really consider the probabilities of the four different cases of restriction site comparisons.

When only one type of restriction enzymes with the same r value is used, the appropriate solutions to the equations are

$$\hat{a} = (m_X + m_Y)/(2m_T), \tag{8}$$

$$\hat{\delta} = [-ln\hat{S}]/r, \tag{9}$$

where

$$\hat{S} = 2m_{XY}/(m_X + m_Y). \tag{10}$$

Equation (9) is identical with the formula obtained by NEI and LI (1979) and KAPLAN and RISKO (1981). The variances [$V(\hat{\delta})$ and $V(\hat{a})$], covariance [Cov($\hat{\delta}$, $\hat{a}$)], and correlation [$r(\hat{\delta}, \hat{a})$] of $\hat{\delta}$ and $\hat{a}$ are given by

$$V(\hat{\delta}) = (2 - S)(1 - S)/(2r^2\bar{m}S), \tag{11}$$

$$V(\hat{a}) = \bar{m}(1 + S - 2a)/(2m_T^2) \approx \bar{m}(1 + S)/(2m_T^2) \tag{12}$$

$$\mathrm{Cov}(\hat{\delta}, \hat{a}) = -(1 - S)/(2rm_T), \tag{13}$$

$$\begin{aligned} r(\hat{\delta}, \hat{a}) &= -[S(1 - S)/\{(2 - S)(1 + S - 2a)\}]^{1/2} \\ &\approx -[S(1 - S)/\{(2 - S)(1 + S)\}]^{1/2}, \end{aligned} \tag{14}$$

whereas the variance of $\hat{S}$ is

$$V(\hat{S}) = S(1 - S)(2 - S)/(2\bar{m}), \tag{15}$$

where $\bar{m}$ is $m_T a$. For estimating the above quantities, we can replace S and $\bar{m}$ by $\hat{S} = 2m_{XY}/(m_X + m_Y)$ and $\hat{m} = (m_X + m_Y)/2$, respectively, in (11) through (15).

Incidentally, if we assume that a is a constant rather than a variable, we still get (9) as an estimate of δ. In this case, however, $V(\hat{\delta})$ becomes $V(\hat{\delta}) = (1 - S)/[r^2\bar{m}S(1 + S)]$. This agrees with KAPLAN and RISKO'S (1981) formula.

It is also approximately the same as NEI and TAJIMA'S (1981) formula, which was derived by considering the variances and covariances of m_X, m_Y and m_{XY}. Their formula is

$$\begin{aligned} V(\hat{\delta}) &= \frac{1}{r^2\bar{m}S}\left[(1 - S) - S(1 - S^{1/2})\left(\frac{3 - S^{1/2}}{2}\right)\right] \\ &= \frac{1 - S}{r^2\bar{m}S(1 + S)} - \frac{(1 - S^{1/2})^4}{2r^2\bar{m}(1 + S)}. \end{aligned} \tag{16}$$

Note that the second term in (16) is negligibly small compared with the first term. The discrepancy between (11) and (16) has occurred, because in the derivation of (16) the number of restriction sites (m_0) for the ancestral DNA sequence from which sequences X and Y were derived was assumed to be constant. Actually, m_0 follows the Poisson distribution (NEI and LI 1979), and if we consider this factor, $V(\hat{\delta})$ is obtained by

$$V(\hat{\delta}) = E_{m0}[V(\hat{\delta}|m_0)] + V_{m0}[E(\hat{\delta}|m_0)], \tag{17}$$

where $V(\hat{\delta}|m_0)$ and $E(\hat{\delta}|m_0)$ are the variance and mean of $\hat{\delta}$ for a given value of m_0, respectively. $E_{m0}(\cdot)$ and $V_{m0}(\cdot)$ stand for the mean and variance of the quantity inside the parentheses with respect to the distribution of m_0. The first term in (17) is equal to (16), whereas the second term is $(1 - S^{1/2})^2/(\bar{m}r^2)$. Therefore, (17) becomes identical with (11). That is, (11) can also be obtained by considering the variances and covariances of m_0, m_X, m_Y, and m_{XY}.

When there are more than one type of enzymes used, we must solve (5a) numerically to estimate δ. The maximum likelihood estimate is given by the solution to

$$\Sigma_i r_i \frac{2m_{XYi} - (m_{Xi} + m_{Yi})e^{-r_i\delta}}{(1 - e^{-r_i\delta})(2 - e^{-r_i\delta})} = 0. \tag{18}$$

In practice, numerical solution of this equation is somewhat cumbersome. The standard scoring method of maximum likelihood estimation does not always give a quick convergence. In the next section we shall present a simple iteration method. The variance of $\hat{\delta}$ obtained from (18) is given by

$$V(\hat{\delta}) = 1/\sum_{i=1}^{k}[1/V(\delta_i)], \tag{19}$$

where $V(\delta_i)$ is the value of (11) for the ith type of enzymes when $S = e^{-r_i\hat{\delta}}$ is used.

In the above formulation we have assumed that $S = e^{-r\delta}$. This assumption is sufficiently accurate as long as $\delta \leq 0.25$. However, if δ is larger than 0.25, it is advisable to use the relationship $S = (1 - \pi)^r$, as mentioned earlier. In this case π can be estimated by solving

$$\Sigma_i r_i \frac{2m_{XYi} - (m_{Xi} + m_{Yi})(1 - \pi)^{r_i}}{[1 - (1 - \pi)^{r_i}][2 - (1 - \pi)^{r_i}]} = 0. \tag{20}$$

If the estimate ($\hat{\pi}$) of π is obtained, $\hat{\delta}$ is given by

$$\hat{\delta} = -\frac{3}{4} ln \left(1 - \frac{4}{3}\hat{\pi}\right). \tag{21}$$

The variance of $\hat{\pi}$ for one type of restriction enzymes is

$$V(\hat{\pi}) = \frac{(1 - \pi)^2(2 - S)(1 - S)}{2r^2\bar{m}S}, \tag{22}$$

whereas the variance of $\hat{\delta}$ and the covariance of $\hat{\delta}$ and $\hat{a}$ are

$$V(\hat{\delta}) = \frac{9(1 - \pi)^2(2 - S)(1 - S)}{2r^2\bar{m}(3 - 4\pi)^2S}, \tag{23}$$

$$\mathrm{Cov}(\hat{\delta}, \hat{a}) = -\frac{3(1 - \pi)(1 - S)}{2rm_T(3 - 4\pi)}. \tag{24}$$

$V(\hat{a})$ and $r(\hat{\delta}, \hat{a})$ are again given by (12) and (14), respectively. When different types of restriction enzymes are used, $V(\hat{\delta})$ can be computed by (19) in the same way as that for the case of $S = e^{-r\delta}$.

At this point, it should be noted that (20) becomes identical with (18) if we replace π by $1 - e^{-\delta}$. Therefore, even if we assume $S = e^{-r\delta}$, $\hat{\delta}$ can be obtained from $\hat{\pi}$ by

$$\hat{\delta} = -ln(1 - \hat{\pi}). \tag{25}$$

NUCLEOTIDE-COUNTING METHOD

An estimate of the number of nucleotide substitutions can also be obtained by counting the number of nucleotide differences between the two sequences compared. We consider shared restriction sites, unique (nonshared) restriction sites, and nonrestriction sites, separately. All nucleotides involved in the shared sites must be identical for the two DNA sequences. In other words, when there are $\Sigma_i m_{XYi}$ shared sites, $\Sigma_i r_i m_{XYi}$ nucleotides involved in the sites are identical for the two sequences. In the case of unique sites there is at least one nucleotide difference per r_i nucleotide sites. The expected proportion of different nucleotides for these sites is therefore $\pi/(1 - S_i)$, where π is the expected proportion of different nucleotides per nucleotide site for the entire DNA sequence. We note that there are $\Sigma_i(m_{Xi} + m_{Yi} - 2m_{XYi})$ unique restriction sites.

The last group of sites is that of nonrestriction sites for both DNA sequences. There are $m_T - \Sigma_i(m_{Xi} + m_{Yi} - m_{XYi})r_i$ nucleotide sites involved in this group, and the expected proportion of different nucleotides is $(1 - \Sigma_i a_i r_i S_i)\pi$. The latter value can be obtained by considering the *expected* numbers of shared sites $[m_T\Sigma_i a_i r_i S_i]$, unique sites $[2m_T\Sigma_i a_i r_i(1 - S_i)]$ and nonrestriction sites $[m_T\{1 - \Sigma_i a_i r_i(2 - S_i)\}]$ (see NEI and LI 1979). That is, the expected number of different nucleotide sites for the entire sequence is

$$m_T\Sigma_i a_i r_i S_i \times 0 + 2m_T\Sigma_i a_i r_i(1 - S_i) \times \frac{\pi}{1 - S_i}$$
$$+ m_T\{1 - \Sigma_i a_i r_i(2 - S_i)\}x = m_T\pi,$$

where x is the expected proportion of different nucleotide sites for nonrestriction sites. If we note $2\Sigma_i a_i r_i \ll 1$, this equation gives

$$x = \frac{1 - 2\Sigma_i a_i r_i}{1 - \Sigma_i a_i r_i(2 - S_i)}\pi$$
$$\approx (1 - \Sigma_i a_i r_i S_i)\pi.$$

We can now determine the proportion of different nucleotide sites in terms of the *observed* numbers of restriction sites. That is,

$$\pi = \frac{1}{m_T}\left[\Sigma_i(m_{Xi} + m_{Yi} - 2m_{XYi})r_i \frac{\pi}{1 - S_i} + \{m_T - \Sigma_i(m_{Xi} + m_{Yi} - m_{XYi})r_i\}(1 - \Sigma_i a_i r_i S_i)\pi\right]. \tag{26}$$

If we replace a_i by $(m_{Xi} + m_{Yi} - m_{XYi})/[m_T(2 - S_i)]$ in (7), this simplifies to

$$\Sigma_i r_i \frac{(m_{Xi} + m_{Yi})S_i - 2m_{XYi}}{(1 - S_i)(2 - S_i)} = 0 \tag{27}$$

approximately. (Note that the expected value of $(m_{Xi} + m_{Yi} - m_{XYi})/[m_T(2 - S_i)]$ is a_i; see NEI and LI 1979.) This is identical with (18) or (20), depending on the definition of S_i. Therefore, the counting method and the maximum likelihood method give the same estimate of δ. This is analogous to the case of estimation of gene frequencies where the gene-counting method and the maximum likelihood method give the same estimate (CEPPELLINI, SINISCALCO and SMITH 1955).

As mentioned earlier, estimation of δ from (18) or (20) is sometimes cumbersome. A simpler method is to use a recurrence formula similar to (26). Formula (26) itself is not very useful for this purpose, because m_T is usually very large compared with the number of restriction sites, and this makes the convergence of π very slow. However, a simple recurrence equation can be obtained from (27). That is, (27) can be written as

$$\Sigma_i \frac{r_i(\hat{m}_i - m_{XYi})}{(1 - S_i)(2 - S_i)} = \Sigma_i \frac{r_i \hat{m}_i}{2 - S_i},$$

where $\hat{m}_i = (m_{Xi} + m_{Yi})/2$. Therefore, we have the following iteration formula.

$$\begin{aligned}\hat{\pi} &= \hat{\pi}_1 \frac{\Sigma_i r_i(\hat{m}_i - m_{XY})/[(1 - S_i)(2 - S_i)]}{\Sigma_i r_i \hat{m}_i/(2 - S_i)} \\ &= \hat{\pi}_1 \frac{\Sigma_i r_i(\hat{m}_i - m_{XYi})/[\{1 - (1 - \hat{\pi}_1)^{r_i}\}\{2 - (1 - \hat{\pi}_1)^{r_i}\}]}{\Sigma_i r_i \hat{m}_i/[2 - (1 - \hat{\pi}_1)^{r_i}]}.\end{aligned} \tag{28}$$

where $\hat{\pi}_1$ is a trial value of $\hat{\pi}$. When $\hat{\pi} = \hat{\pi}_1$, $\hat{\pi}$ is the maximum likelihood estimate of π. In practice, we can first estimate π by using (9) or $\hat{\pi}_1 = 1 - S^{1/r}$ for a particular kind of restriction enzymes and use it in (28). Usually, four or five cycles of iterations are sufficient for getting the maximum likelihood estimate. A computer program for computing $\hat{\delta}$ and its standard error is available upon request.

VALUES OF a AND r FOR RESTRICTION ENZYMES WITH MULTIPLE RECOGNITION SEQUENCES

Although most restriction enzymes recognize a unique sequence of 4, 5, or 6 nucleotides, others recognize multiple sequences of a given number of nu-

cleotides. For example, *Hind*II recognizes the sequence GTPyPuAC, where Py is either T or C and Pu is either A or G. In this case a is given by $(g_1 + g_2)(g_3 + g_4)g_1g_2g_3g_4$, whereas $S = (1 - \pi)^4(1 - 2\pi/3)^2$. S can be approximately written as $(1 - \pi)^{16/3}$ for $\pi \leq 0.3$. Therefore, if we redefine r as $r = 16/3$, $S = (1 - \pi)^r$ still holds (NEI and TAJIMA 1981). Another 6-base enzyme with multiple recognition sequences is *Hae*I, which recognizes $\binom{A}{T}$GGCC$\binom{A}{T}$. In this case $a = (g_1 + g_2)^2g_2^2g_4^2$, and $S \approx (1 - \pi)^{16/3}$. Namely, r is the same as that for *Hind*II. The values of a and r for nine different types of restriction enzymes are given in Table 1. Most restriction enzymes (type II) belong to one of these types, so that we can compute a and r.

NUMERICAL EXAMPLE

GOTOH *et al.* (1979) have compiled restriction sites data for the rat and mouse mitochondrial DNAs (mtDNAs). In the comparison of two strains of the rat, A and B, four six-base enzymes, one four-base enzyme and two multiple-sequence enzymes with $r = 16/3$ were used. The numbers of restriction sites identified are given in Table 2. We first compute $\hat{\pi}_1$ by using the data for six-base enzymes, where $\hat{S} = 22/23$. $\hat{\pi}_1$ then becomes $1 - (22/23)^{1/6} = 0.0073813$. If we use (28), we have $\hat{\pi}_2 = 0.0097470$ as the second estimate.

TABLE 1

Examples of various types of type II restriction enzymes, expected frequencies of restriction sites (a), and the r *values.*[a]

Enzyme	Recognition sequence	a	r
4-base			
*Hae*III	GGCC	$g_2^2g_4^2$	4
*Mnl*I[b]	CCTC	$g_1g_2^3 + g_3g_4^3$	4
5-base			
*Hin*fI	GANTC	$g_1g_2g_3g_4$	4
*Eco*RII	CC$\binom{A}{T}$GG	$(g_1 + g_3)g_2^2g_4^2$	14/3
*Mbo*II[b]	GAAGA	$g_3^3g_4^2 + g_1^3g_2^2$	5
6-base			
*Eco*RI	GAATTC	$g_1^2g_2g_3^2g_4$	6
*Hae*I	$\binom{A}{T}$GGCC$\binom{A}{T}$	$(g_1 + g_3)^2g_2^2g_4^2$	16/3
*Hind*II	GTPyPuAC	$(g_1 + g_2)(g_3 + g_4)g_1g_2g_3g_4$	16/3
7-base			
*Eco*I	GGTNACC	$g_1g_2^2g_3g_4^2$	6

[a] One example from each type is given.

[b] *Mnl*I recognizes double strand sequences $\begin{matrix}CCTC\\GGAG\end{matrix}$ and $\begin{matrix}GAGG\\CTCC\end{matrix}$, whereas *Mbo*II recognizes $\begin{matrix}GAAGA\\CTTCT\end{matrix}$ and $\begin{matrix}TCTTC\\AGAAG\end{matrix}$. However, these enzymes are used very rarely.

TABLE 2

Numbers of restriction sites observed in the comparisons of mitochondrial DNA between rat strains A and B and between the rat and mouse

Enzymes used	r	$\frac{m_X + m_Y}{2}$	m_{XY}
Rat (A)-Rat (B)			
Six-base[a]	6	23	22
Four-base[b]	4	22	21
Multiple-sequence[c]	16/3	6.5	6
Rat-Mouse			
Six-base[d]	6	16.5	3
Four-base[e]	4	7	2
Multiple-sequence[f]	16/3	4	1

[a] *Bam*HI, *Eco*RI, *Hind*III, *Hpa*II.
[b] *Hae*III.
[c] *Hae*II, *Hind*II.
[d] *Bam*HI, *Eco*RI, *Hind*III, *Hpa*I, *Pst*I.
[e] *Hha*I.
[f] *Hae*II, *Hind*II.

Further iterations give $\hat{\pi}_3 = 0.0097983$ and $\hat{\pi}_4 = 0.0097983$. We can therefore take $\hat{\pi} = 0.00980$ as the maximum likelihood estimate. This is slightly smaller than GOTOH *et al.*'s estimate (0.0103). If we use (19), (21) and (23), we have $\hat{\delta} = 0.00986 \pm 0.00453$, whereas if we use (11), (19) and (25), $\hat{\delta} = 0.00985 \pm 0.00451$. The difference between the two estimates of δ is negligible.

A similar set of restriction-site data exists for the comparison of the rat and mouse (Table 2). In this case the data for six-base enzymes give $\hat{\pi}_1 = 0.247327$, and after a few cycles of iterations we have $\hat{\pi} = 0.250$, which is slightly larger than GOTOH *et al.*'s estimate (0.244). Equation (21) gives $\hat{\delta} = 0.303 \pm 0.073$, whereas equation (25) gives $\hat{\delta} = 0.287 \pm 0.065$. Therefore, the difference between the two estimates is appreciable in this case.

In this connection it should be mentioned that, although our method gave $\hat{\pi}$ values similar to GOTOH *et al.*'s in the present case, the two methods do not always give similar values, particularly when S_i is close to 0.5. Furthermore, GOTOH *et al.*'s formulation seems to give an erroneous value of variance, as will be seen from the comparison of their equation (10a) and our equation (5a).

DISCUSSION

In the present paper we have assumed that the four nucleotides T, C, A, and G are randomly arranged in the DNA sequence under investigation. In practice, this assumption usually does not hold. ADAMS and ROTHMAN (1982) recently reported that the distribution of restriction sites is significantly nonrandom in the human mtDNA. They claimed that this would introduce a serious error in the estimate of nucleotide differences obtained by the restriction enzyme method. However, their study is based on the assumption that

the mutation rate is the same for all nucleotide pairs and thus the expected frequencies of the four nucleotides are equal. In practice, this assumption does not hold in the mtDNAs of most organisms (BROWN 1981; BROWN *et al.* 1982; AQUADRO and GREENBERG 1983), and furthermore, it is not really necessary for our purpose as long as different types of nucleotides are arranged at random. Even if the rate of nucleotide substitution varies with nucleotide pair, the effect of the variation on the average number of nucleotide substitutions is generally small as long as $\delta < 0.5$ (TAKAHATA and KIMURA 1981; GOJOBORI, ISHII and NEI 1982). In this case, λ in (2) should be interpreted as the average rate of nucleotide substitution per site as defined by KIMURA (1981) and GOJOBORI, ISHII and NEI (1982). In the case of mitochondrial DNA, transitional nucleotide substitutions are much more frequent than transversional changes (BROWN *et al.* 1982; AQUADRO and GREENBERG 1983). However, even this extreme type of nucleotide substitution does not seem to affect the estimate of δ seriously unless δ is larger than 0.3 (M. NEI and F. TAJIMA, unpublished results). We note that the restriction enzyme technique is not usually used when $\delta > 0.3$, since $V(\hat{\delta})$ is very large in this case (LI 1981).

Of course, when the rate of nucleotide substitution is not the same for all nucleotide pairs and a certain type of restriction enzymes are used, some bias is expected to occur in the estimate of δ (AOKI, TATENO and TAKAHATA 1981; TAJIMA and NEI 1982). For example, if the substitution rate between nucleotides A and T is much higher than that for the other pairs of nucleotides, the values obtained from the enzyme with recognition sequence GGCC (*Hae*III) or GCGC (*Hha*I) would be underestimates. However, TAJIMA and NEI (1982) have shown that this type of bias is generally very small if many different kinds of restriction enzymes are used.

In the present paper we have been concerned with the estimation of nucleotide substitutions between a pair of DNA sequences. Many biologists are, however, interested in estimating a phylogenetic tree of organisms rather than DNA sequences, and each oganism is usually polymorphic. In this case we must sample several DNA sequences from each organism (or population), and δ should be estimated by taking into account the intrapopulational variation, as shown by NEI and LI (1979). To reduce the sampling error of $\hat{\delta}$, however, it is important to use a large number of restriction enzymes. As long as the number of restriction enzymes used is large, the number of DNA sequences sampled from a species can be relatively small (F. TAJIMA, unpublished data).

We thank ARAVINDA CHAKRAVARTI and NORMAN KAPLAN for their comments on an earlier version of this paper. This work was supported by research grants from the National Institutes of Health and the National Science Foundation.

LITERATURE CITED

ADAMS, J. and E. D. ROTHMAN, 1982 Estimation of phylogenetic relationships from DNA restriction patterns and selection of endonuclease cleavage sites. Proc. Natl. Acad. Sci. USA **79:** 3560–3564.

AOKI, K., Y. TATENO and N. TAKAHATA, 1981 Estimating evolutionary distance from restriction maps of mitochondrial DNA with arbitrary G + C content. J. Mol. Evol. **18:** 1–18.

AQUADRO, C. F. and B. D. GREENBERG, 1983 Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals. Genetics **103:** 287–312.

BROWN, W. M., 1981 Mechanisms of evolution in animal mitochondrial DNA. Ann. N.Y. Acad. Sci. **361:** 119–134.

BROWN, W. M., E. M. PRAGER, A. WANG and A. C. WILSON, 1982 Mitochondrial DNA sequences of primates: tempo and mode of evolution. J. Mol. Evol. **18:** 225–239.

CEPPELLINI, R., M. SINISCALCO and C. A. B. SMITH, 1955 The estimation of gene frequencies in a random-mating population. Ann. Hum. Genet. **20:** 97–115.

GOJOBORI, T., K. ISHII and M. NEI, 1982 Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. J. Mol. Evol. **18:** 414–423.

GOTOH, O., J. I. HAYASHI, H. YONEKAWA and Y. TAGASHIRA, 1979 An improved method for estimating sequence divergence between related DNAs from changes in restriction endonuclease cleavage sites. J. Mol. Evol. **14:** 301–310.

JUKES, T. H. and C. R. CANTOR, 1969 Evolution of protein molecules. pp. 21–123. In: *Mammalian Protein Metabolism*. Edited by H. N. MUNRO. Academic Press, New York.

KAPLAN, N. AND C. H. LANGLEY, 1979 A new estimate of sequence divergence of mitochondrial DNA using restriction endonuclease mapping. J. Mol. Evol. **13:** 295–304.

KAPLAN, N. and K. RISKO, 1981 An improved method for estimating sequence divergence of DNA using restriction endonuclease mappings. J. Mol. Evol. **17:** 156–162.

KIMURA, M., 1981 Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. USA **78:** 454–458.

LI, W.-H., 1981 A simulation study on Nei and Li's model for estimating DNA divergence from restriction enzyme maps. J. Mol. Evol. **17:** 251–255.

NEI, M. and W.-H. LI, 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA **76:** 5269–5273.

NEI, M. and F. TAJIMA, 1981 DNA polymorphism detectable by restriction endonucleases. Genetics **97:** 145–163.

TAJIMA, F. and M. NEI, 1982 Biases of the estimates of DNA divergence obtained by the restriction enzyme technique. J. Mol. Evol. **18:** 115–120.

TAKAHATA, N. and M. KIMURA, 1981 A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. Genetics **98:** 641–657.

UPHOLT, W. B., 1977 Estimation of DNA sequence divergence from comparison of restriction endonuclease digests. Nucleic Acids Res. **4:** 1257–1265.

Corresponding editor: C. F. Wehrhahn